AD/A-003 483

INFORMATION PROCESSING ANALYSIS OF VISUAL PERCEPTION: A REVIEW

A. J. Thomas, et al

Stanford University

Prepared for:

Advanced Research Projects Agency

June 1974

STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO AIM-227

JUNE, 1974

COMPUTER SCIENCE DEPARTMENT
REPORT NO. CS-408

INFORMATION PROCESSING ANALYSIS OF VISUAL PERCEPTION

A REVIEW

by
A.J.Thomas and T.O.Binford

Abstract:

We suggest that recent advances in the construction of artificial vision systems provide the beginnings of a framework for an information processing analysis of human visual perception. We review some pertinent investigations which have appeared in the psychological literature, and discuss what we think to be some of the salient and potentially useful theoretical concepts which have resulted from the attempts to build computer vision systems. Finally we try to integrate these two sources of ideas to suggest some desirable structural and behavioural concepts which apply to both the natural and artificial systems.

The writing of this paper, and some of the research described herein, was supported by the Advanced Research Projects Agency of the Office of the Secretary of Defense under Contract No. DAHC 15-73-c-0435.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S.Government.

Reproduced in the USA. Available from the National Technical Information Service, Springfield, Virginia 22151.

I. INTRODUCTION.

In recent years there has been a greatly increased level of conversation between students of psychology, and of artificial intelligence(AI). This increase seems to stem, on the one hand, from AI's partial acquiescence to the notion that psychological evidence, particularly as couched in the models due to Neisser(Neisser,1967), Norman(Norman,1970) and others, may have a place in its own thinking about cognitive processes, and on the other, from a spreading appreciation on the part of psychologists of the potential virtues of the computational metaphor. This latter movement has been particularly evident in the literature on memory structures(e.g.Anderson & Bower,1973), but seems as yet not to have had significant impact upon studies of the perceptual process. That the study of the mechanisms of perception of real-world scenes (which we sharply distinguish from the reading process) is not widespread in experimental psychology at the present moment seems partly to be a matter of fashion and, more importantly, to be due to the apparent unavailability of powerful information-processing concepts on a par with those concerning the organization of memory.

In this essay, we wish to suggest that recent advances in the construction of artificial vision systems provide some pointers to such an information-processing theory, and that the time is ripe for an effort to integrate computational ideas and empirical investigation. In attempting to further this suggestion, we outline what we think to be some of the salient and potentially fruitful concepts which AI has generated; we review what we take to be some pertinent investigations which have

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER STAN-CS-74-408	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER AD A-003483
4. TITLE (and Subtitle) INFORMATION PROCESSING ANALYSIS OF VISUAL PERCEPTION: A REVIEW.		5. TYPE OF REPORT & PERIOD COVERED technical, June 1974
		6. PERFORMING ORG. REPORT NUMBER STAN-CS-74-408
7. AUTHOR(s) A. J. Thomas and T. O. Binford		8. CONTRACT OR GRANT NUMBER(s) DAHC-15-73-c-0435
9. PERFORMING ORGANIZATION NAME AND ADDRESS Computer Science Department Stanford University Stanford, California 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS ARPA/IPT, Attn: Stephen D. Crocker 1400 Wilson Blvd., Arlington, Va. 22209		12. REPORT DATE June, 1974 13. NUMBER OF PAGES 60
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) ONR Representative: Philip Surra Durand Aeronautics Bldg., Rm. 165 Stanford University Stanford, California 94305		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Releasable without limitations on dissemination.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We suggest that recent advances in the construction of artificial vision systems provide the beginnings of a framework for an information processing analysis of human visual perception. We review some pertinent investigations which have appeared in the psychological literature, and discuss what we think to be some of the salient and potentially useful theoretical concepts which have resulted from the attempts to build computer vision systems. Finally we try to integrate these two sources of ideas to suggest some desirable structural and behavioural concepts which apply to both the natural and artificial systems.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

appeared in the psychological literature; and we attempt to gather these two threads together to begin weaving the fabric of a coherent information-processing approach to the perception of the visual world.

It would be wise at this point to make clear our concern over the problem raised by the different hardware structures available to artificial and natural vision systems: our arguments will stress commonality of computational processes, rather than any features of the implementation of these processes, whether in a digital computer or the brain. We think it worthwhile, however, to devote some effort to suggesting how processing which may be essentially serial in an artificial system may have its nature altered when it is considered in the context of the (at least partly) parallel mechanisms available to the brain.

2. DEFINING THE PROBLEMS.

In this section we attempt to set out what we think to be the major problems to whose solution an information-processing theory of perception should aspire. A large number of these problems stem from a realisation that a large part of perception is *not* concerned with *recognizing* objects which may already have appeared in the repertoire of our experience, but rather with the *description* of scenes. That is to say, the problem is that of transforming a large body of sensory data, for example, intensity, colour and texture at a large set of points within the visual field, into a more compact description or representation of the scene being viewed, so that this description may be incorporated into, and interact with, the large corpus of information, both visual and non-visual, which may already be stored in memory. Our claim is that this view of the nature of perception, uncontroversial though it may seem at first glance, immediately reduces the usefulness of a large body of techniques, historically labelled 'pattern classification', and largely based on template matching, which were developed for recognition of simple, planar geometric patterns such as alphanumeric characters. Furthermore, it suggests the necessity for representational formalisms which capture in some smooth way the whole range of sensory and symbolic information to which humans have access.

Several questions must immediately be asked of this data-description process:

(a) just which perceptual constructs are abstracted from the sensory data;

(b) what is the computational structure of this process of abstraction which gives rise to these constructs;

(c) how long are the retinal and edge data, etc., stored? Are they kept in short-term memory during some verification process, or for calculation of motion parallax;

(c) are the data discarded at all, in fact, or do we remember images for long periods? If not, just what form does the representation in long-term memory take which allows the reconstruction of the mental images with which we are subjectively familiar? Is the representation purely pictorial in nature, or purely symbolic, or does it have some of the characteristics of both? We conjecture that at the basic level of representation, there is no sense in making a distinction between the pictorial and symbolic. When we say 'pictorial' we do not mean to suggest that the representation is isomorphic to a retinal image; rather, that the representation is a graph-like structure whose nodes represent pictorial elements (e.g. edges, vertices, regions, volumes, &c.). These nodes may have verbal labels associated with them.

We approach our task initially via a fairly detailed examination of some possible stages in the processing of visual information: what kinds of data are needed at the various stages?; what kinds of descriptive constructs can each stage generate on its own?; what help is needed from higher level processes? Our suggestion is that a strictly hierarchical structure, such as might be inferred for instance from the original ideas of Hubel and Wiesel(Hubel & Wiesel,1968), is not appropriate for the task in hand. A visual system which consists of a discrete set of analytical stages, each encoding its own specialized conclusions for communication to an immediate superior, cannot, we argue, perform the kinds of computations necessary for the understanding

of visual input. The alternative structure is *heterarchical*, which is to say that there is no strict linear ordering of computational steps. In a heterarchical system, information flow can be bi-directional - advice and queries from more proximal stages can influence the behaviour of more distal stages. We might note in passing that such an organization seems to be implied by several neurophysiological studies which have recently come to hand. Horn and Hill(Horn & Hill,1969) demonstrated gravitational (presumably vestibular) effects on the properties of visual orientation detectors in cat striate cortex; Spinelli and his colleagues (Spinelli,Starr & Barrett,1968), in a series of studies, have demonstrated an influence of auditory and somaesthetic stimuli on the shapes of retinal receptive fields; and other studies have shown changes in retinal receptive field sizes with changing activity in the oculo-motor system (accomodation). Such effects are consonant with the idea that context effects can alter the behaviour of even the most elementary stages of perceptual analysis.

We should also take note here that several recent experimental studies have suggested that the hierarchical mechanisms proposed by Hubel and Wiesel to account for the apparent progressive specialization of properties amongst visual cortical cells (from simple to complex to hypercomplex) may not be correct. Hoffman and Stone(Hoffman & Stone,1971;Hoffman,1973) in a study of correlations between receptive-field properties and conduction velocities of such cells found, firstly, that at least 40 percent of complex cells are activated monosynaptically by fast afferent fibers with delay times of 1-2 msec., while simple and hyper-complex cells are not activated mono-synaptically by fast afferents, but some proportion is activated mono-

synaptically by slow fibers (3-5 msec.). Secondly, they have shown that fast and slow fiber activity is relayed separately in the lateral geniculate nucleus. They have suggested that simple, complex and hypercomplex cells process visual information in parallel rather than in the serial manner proposed earlier.

We actually want to make the strong claim that visual analysis of scenes such as we see in every-day life is much more difficult than one might at first assume, and that the very richness of the information available to us imposes significant constraints upon the nature of an analytical system which is able to cope with it. Consider some of the simplest problems:

(a) the volume of data: at present the Hand-Eye project at Stanford uses for its visual sensor a television camera which generates data consisting of a four-bit (i.e. 16 grey-level) description of light intensity at each of about 330 by 260 points. The camera generates these data at about one scan every 60th. of a second,but it takes far longer to analyse it even to the fairly primitive level that we have achieved so far. The application of an edge-operator to each of these 80,000 points takes a minimum of 150 microseconds/point on our present machine. When one realizes further that this volume of analysis may have to be done on each of several stereo pictures (for depth by correlation), with colour, it is easy to see that processing time can get out of hand quite rapidly. Compare this with the human visual system, which has about an order of magnitude more resolving power, and yet manages to do much more in a small fraction of a second. Naturally, the human system has much more

specialized hardware, but we hope that the point will emerge that even so the data volume problem is considerable;

(b) noise: the human eye is nothing like a perfect camera, but has considerable optical and electrical defects, anisotropies in resolution and abberation effects;

(c) the eyes are not fixed: they move both with the rest of the body frame and independently in their orbits. Why is it then that the visual world appears to be stabilised under these movements? Helmholtz (Helmholtz,1963) pointed out that passive movement of the eyeball causes a disconcerting shift in the perceived field, and Brindley and Merton(Brindley & Merton, 1960) showed that the apparent stability is not due to feedback from proprioceptors in the orbital muscles themselves, but rather to correction for the movement at a much higher level, involving direct interaction of the oculo-motor and visual centers;

(d) variable view point: three-dimensional objects, unlike characters on a piece of paper, can present many and varied appearances. Because of the fact that we are separated from them in three-space, distortions are caused by the phenomenon of *perspective*. If we are are going to design a visual system , we have to have some fairly sophisticated way of recognising objects from various viewpoints, and even making predictions about the appearances of parts of objects that we can't see. That is to say, we must have an efficient way of mapping the visual appearance onto descriptions which we already have stored. We have to develop ways of dealing with transformations which are translational and rotational, and simultaneously deal with scale changes as we approach or recede

from objects. For an approach based on matching features from templates in any simple-minded way, in a three-dimensional world, the obscuration of one object by another yields yet another fatal difficulty. Just how difficult and subtle problems of three-dimensional visual geometry are is suggested by the fact that it was not until the Renaissance that artists (da Vinci,1956;Alberti,1547) were able to formulate rules for the naturalistic representation of space in painting. Painting and art in general raise another issue, which is the inverse of the problem of everyday scenes - that of the cartoon. Why is it that so much visual information can be conveyed by a few strokes of the brush or pen? What are the mechanisms which allow us to fill in? Art reveals most clearly the issue of interpretation of pictures (see for example the excellent discussion by Gombrich(Gombrich,1960). Remember the duck/rabbit cartoon (Wittgenstein,1953), and Boring's Young Woman/Old Hag (Boring,1942) drawing, as well as examples of figure-ground confusions which are all aspects of the very difficult question of multi-stable states in vision, which crops up both in depth-perception and in interpreting line drawings as three dimensional objects. Why does the Necker Cube apparently have *two* stable three-dimensional interpretations, as well as the perfectly simple two-dimensional one? What are the factors which decide which of the stable states we land in, and why is it so difficult to switch from one state to another?

Actually we hope that it will become clear later that the three-dimensionality of the world is a help rather than a hindrance to scene analysis. Techniques have been developed(Nevatia & Binford, 1973) for using depth information alone

for the segmentation of scenes containing curved objects and these same workers have proposed a volume-based representation for such objects. We discuss this in more detail in Section 5;

(e) variability in lighting: light can be of many colours and intensity, and come from one or more sources, as well as being reflected off some surfaces onto adjacent ones. This again is a source both of annoyance and information. It's annoying that one has to keep changing the characteristics of one's sensory mechanism all the time - one has to adapt for varying light levels. On the other hand, if one knows something about the lighting situation one can use information from shadows and shading to reconstruct some aspects of the shapes of objects. A later section will discuss some of the problems which arise in judging the colours of object under these kinds of difficulties. That too will be an example where there are some quite good information processsing models which are borne out by psychophysical data.

Now that we have laid out some of the difficulties, we should comment on what we take to be the correct and incorrect methods for coping with them. The basic thrust of our argument will be that naive, brute-force methods such as those developed in the field of pattern classification for recognising printing, hand-writing, bubble-chamber photographs and so on just will not do at all for analysing the kinds of scenes with which humans are normally confronted. There seem to be some very striking parallels here with attempts at machine understanding of natural language. At the beginning of the sixties, people in that field, and linguists of the MIT School, e.g. Chomsky, Fodor, Katz and their colleagues, thought that the way to

describe natural language understanding systems (whether machine or human) was to write sophisticated context sensitive generative grammars of one kind or another.It soon became clear that this purely syntactic way of going about things did not seem to work. Language is not wholly a system of rewriting rules applied to base symbols to produce terminal strings. It is instead a system for conveying information. And information is not measured in terms of bytes: its significance (its content) is very dependent upon the circumstances surrounding its utterance and reception. It is dependent upon a wider kind of context than the transformational grammarians had in mind - the context is the state of the world pertaining at the time somebody says something; and this state of affairs can involve all kinds of non-linguistic factors. So recently there has been a much more serious effort to develop truly *understanding* language processors (e.g.Winograd,1973). Exactly this kind of historical pattern has been apparent in the psychology of visual perception, but over a longer time scale. It may seem paradoxical in view of their well known antipathy that, loosely speaking, the late 19th. century behaviourists were akin in their spirit to the generative syntacticists, in stressing the purely atomistic structural features of their respective phenomena . The Gestaltist reaction went to the opposite extreme of stressing the wholistic features of perception to the detriment of the possibility of partitioning the perceptual process into pieces suitable for empirical investigation.

Before going further, we must clarify our notion of 'level of analysis'. We have already introduced the notion of a heterarchical system without defining what we mean by a 'level' or 'stage' within such a system. There seem to be two useful ways of formulating such a definition:

the first involves a dependence on the type of data which a process accepts as its input. According to such a criterion, processes dealing with raw intensity data (retinal receptors,bipolars, etc.)are at the very lowest level; those which deal with edges are at a higher level, but are themselves subordinate to processes which take more global input (from several regions of the visual field, for example);

the second way depends upon the amount of 'advice' (i.e.interaction from processes at its own or a higher level) that a process needs in order successfully to accomplish its task. By this criterion, the lateral inhibition and Retinex processes that we shall shortly describe are at a different level from, for instance, a region-association process.(In fact, for consistency, we rate the level derived by this criterion as inversely proportional to the advice it needs).

In a heterarchical system, as we have described it, the clean notion of level we have just laid out becomes clouded; a better measure is perhaps the informational connectivity of a process, i.e. a measure on the number of sources from which it derives advice and instruction. Crudely speaking, the less advice that a process needs, the more likely it is to have an efficient implementation in a parallel, as opposed to a serial, system. Of course, advice may itself be treated as data.

3. AN ANALYSIS OF LOW-LEVEL MECHANISMS.

In this section, we will look in detail at the kinds of information processing which might occur at some early stages in the visual pathway; in particular, we ask of each of these stages what it can achieve on its own, with no guidance from other processes occurring at the same or higher levels (using 'levels' in the rather careful sense that was laid out in the previous section). This analysis will also reflect upon the arguments concerning serialism vs parallelism which we brought up in that section. We may find that there are some processes which can apparently be carried out completely at one level (and presumably implemented efficiently in parallel hardware); while, in looking at other processes, we will conclude that they cannot function without information from other stages, and perhaps in the last instance without non-visual information about the nature of the world.

(a) Edges.

Neurophysiological studies suggest that cells in the mammalian visual cortex extract edges from retinal data (e.g. Hubel & Wiesel,1968). However, edges between regions of uniform intensity are only one useful perceptual construct. The surfaces of hair, leaves, feathers and foam have irregular texture, as do other surfaces. Gibson has of course argued that textures and gradients of texture are important perceptual constructs (Gibson,1950). AI has only recently begun to deal with textured scenes; these studies are discussed in Section 3c. Since our

perceptual goals are interpretations in terms of objects and their spatial relations, stereo and motion parallax depth perception can also provide information most closely related to these perceptual goals. These abilities are apparently developed at birth (Bower,). Perhaps other visual learning depends on the segmentation structure provided by depth perception. AI studies on depth are discussed in Section 3d.

Natural scenes usually contain objects with well-defined surface boundaries. The retinal projections of such scenes are patchworks of areas of obscuring objects. Some edges correspond to boundaries between objects, some to interior edges of objects, others to surface markings, others to shadows and reflections. All signify features of potential interest, but there is no direct connection between intensity boundaries and the spatial interpretations which are the perceptual goal. Thus, it is possible to find intensity discontinuities on a local basis, even though some edges with low contrast will be missed. It is not possible to find the "meaningful edges" on a local basis, i.e. to make a "perfect line drawing". (This one reason why recent successes in the development of higher level segmentation algorithms have not made the perceptual problem redundant - all have been dependent on being provided with prefect line drawings as input.) Present techniques of finding intensity discontinuities require extensive calculation and are not really adequate. We expect that performance approaching human visual acuity inherently can only be achieved at great computational cost (e.g. the retina has 1000 times as many cells as our TV images have resolution elements), and that substantial economies in

computation are possible only using sampling strategies which ignore much of the image. Thus, sampling coarsely allows detection of extended boundaries, from which many finer details (but not all) can be determined. For many problems of interest, however, those reduced abilities are adequate. In future, vision analysis programs will have to devote much attention to strategies which take account of extension in space or continuity in time, or for which context greatly limits the resolution necessary.

In this section we prefer to discuss boundary finding (for regions without texture) in terms of computational components: local edge operators and edge organization procedures. In the biological system, the local edge operators are Hubel-Wiesel cells, defined over small disks (typically 1 degree of arc in the monkey). These cells are strongly directional, having an angular resolution of 5 degrees. There are approximately 20,000 such cells in the striate cortex. A computational equivalent is the Hueckel operator (Hueckel, 1971), defined over a small disk. This operator is directional in that it calculates moments of the intensity function in a few directions and determines the direction of best fit of a single edge on the disk. We can estimate the computational cost of sampling a picture coarsely with this operator. We want to detect edges at all orientations. Since our initial concern is with features which are considerably spatially extended we can get away with sampling at only a few positions along edges, but must sample at many positions perpendicular to an edge. It is enough to apply the Hueckel operator at every other point along widely separated vertical and horizontal lines. Let us say

we use 8 horizontal and 8 vertical lines. At about 7 msec. per application, this amounts to 15 seconds for a 256 by 256 image, or 90 seconds in three colors with stereo views. This is about a factor of 20,000 slower than comparable human operations, for much coarser resolution (about a 1000 times fewer points).

What should be the design criteria for a local edge operator? Such an operator describes the light intensity surface over a small disk by a step function normal to a line through the disk. Its utility is measured by its sensitivity and computational cost for a specified error rate (e.g. sensitivity defined by the contrast required for 50% positive responses at an edge, for a threshold set for 5% false positives where there is no edge). We can model light intensity at an edge by the edge signal with detector noise added. Detector noise may be confounded with surface specks and markings. The former may be readily modelled on the basis of electrical or electro-physiological analysis but the latter is basically a signal which may or may not be significant; the contrast of such specks may be arbitrarily large: to attempt to eliminate response to them by simply raising thresholds would lower sensitivity. They are systematic phenomena on a local scale, but are characterised by the fact that they are not extended. It must be left to the edge organization process to deal with such specks.

There is a straight-forward tradeoff between sensitivity on the one hand, and computation cost and resolution on the other. This suggests the use of a range of sizes of operators, with sizes covering the spectrum from the dimensions of specks to those for extended edges. Unfortunately, since the Hueckel operator is

sensitive primarily to the intensity gradient, non-uniform illumination, predominantly from reflections, can cause it to return false positive results over an entire surface. The biological system is known to be insensitive to smooth gradients; lateral inhibition serves to eliminate these effects. An edge operator using this technique has been designed by Horn and Binford (Horn & Binford, 1973). A valuable elucidation of its possible manifestation is given by Horn and by Marr (Horn, 1974; Marr, 1974)

To what extent is improvement in sensitivity possible for local edge operators? The Hueckel operator requires a threshold only 1.5 times optimum (we assume that the difference of means across an edge is nearly optimal). If we assume that two multiply-add operations per point is minimal, the Hueckel operator is about a factor of 4 slower than optimal. There seems little room for dramatic improvements in the computational behaviour of such edge operators. Finding edge fragments is local in the sense that the support of the calculation is a small disk.

One operation of *edge organization* is also local: that of finding edge fragments which are near to each other and have similar slopes. Little is known from physiology about mechanisms for the organization of edge information from Hubel-Wiesel cells. One conjecture is that their output goes directly to form a Fourier transform. We argue later against the utility of that conjecture. There are many models from machine perception. One of these is the boundary of a connected region; another is the edge-following technique; the edge at each point contains directionality information which enables an operator to be applied only in regions which are predicted to have edges traversing them.

Local edge operators are useful for curves which are locally straight over the disk of the operator. For a large disks, the intensity surface will usually be more complex than a step-function. There are many more possibilities in the case of several lines, of curves, or of small textural features. One way of coping is to find optimal curves which maximize some local contrast function (e.g. the gradient along the curve). That approach is very expensive computationally because of the enormous number of possible enumerations of adjacent points, but by neatly enumerating combinatorics and by use of continuity, the expected shape of the curve, and heuristic search, these techniques have been made feasible in situations such as analysis of x-rays, in which the shape of the curve is known in advance.

Combinatorics can be minimized by a two step process of local edge hypothesis by thresholding the output of a local edge operator followed by edge organization. We assume that such stages follow the Hubel-Wiesel operators in biological systems. There are two global techniques which work well for straight lines: clustering of edge fragments in the space of line parameters (angle and minimum distance) (Perkins & Binford, 1973); and local clustering of edge fragments projected along a variety of directions. For curves, no economical equivalent has been found.

Other techniques are local; the simplest of these is the *region growing* form of edge organization. Clearly, the boundary of the set of points which all satisfy a region-predicate is an edge. Thus region growing is a particular simple form of edge organisation, which lacks a notion of smoothness which would serve to bridge gaps. The aim of the edge organization process is to link edge fragments which are nearby

and have similar slopes. This may be done in an edge following mode (Tenenbaum & Pingle, 1971), in which the edge operator is used to track along edges, or in its parallel equivalent, when the edge operator is applied in a raster scan; edge fragments as they are found are linked to one of a number of nearby unterminated edges. It appears to be difficult to extend the computation of a region uniform in intensity to regions uniform in the spatial distribution of features. We feel that these difficulties are related in part to the inherent computational complexity associated with two-dimensional geometry.

(b) Colour

Given a receptor mechanism sensitive at three conveniently spaced wavelengths, such as is available in the retina, one might think that it would be a relatively easy matter to arrive at judgements of the colours of points or regions within a scene. Unfortunately this is not the case, because the mapping from physical stimulus space (i.e. that containing wavelength and intensity) to sensation space is not an isomorphism. Two operations which are somewhat different are colour matching on the one hand, and colour naming on the other. The former is a resolution problem, while the latter involves the important phenomenon of *colour constancy*.

One of the most elegant demonstrations that the relationship between flux at various wavelengths of reflected light and the associated colour sensations is not straightforward was given by Edwin Land in his William James Lectures at

Harvard (see Land & McCann,1971). His experimental subject was a pastel drawing of a street scene by Jeanne Benton, which included a green awning on one side and a red door on the other. Light at a long wavelength (650nm) was shone onto the awning, while light of a middle wavelength (540nm) was shone onto the door in such a way that the same long wave flux came from the centre of the awning and the centre of the door, and similarly for the medium wavelength. In other words, the total flux incident on the retina from the two regions was identical in intensity and wavelength mixture. In long wavelength light alone the door appeared very light and the awning almost black, and the reverse was true in middle wavelength light alone. However in the mixture of light, the awning appeared green and the door appeared red! Land in his Retinex theory (*op.cit.*) proposed that the colour of objects is determined by their *lightness* computed at three distinct wavelengths. The lightness of a region is an estimate of its reflectance at given wavelength after high-pass filtering has been performed to remove slow changes in incident flux. This is based on the intuition that changes in reflectance are abrupt (at object boundarise) while illumination changes are more gradual. Thus areas that are lighter in long wavelength and dark in middle wavelength light *always* look red, independently of the actual wavelength distribution in the reflected light. Areas that look light at medium wavelengths and dark in long always are perceived as being green.

The Retinex operation allows a deemphasising of shadows and gradual brightness changes across uniform coloured regions, and also corrects for

colour casts, since the colour of any one region is not judged absolutely, but relative to the perceived colours of its near neighbours. A program has been developed by one of us which simulates this model of colour analysis (Thomas,1974). It looks at the scene through three filters, and uses the lightness data at these wavelengths to construct a colour triangle: the extremes of intensity at each wavelength define the vertices and the white point. This program has had some success in dealing with lighting situations involving colored shadows and other difficult casts, and has been found to pass at least some standard tests for anomalous color vision.

Horn (Horn,1974) and Marr (Marr,1974) have investigated the way in which the Retinex operation which we described may be implemented in artificial and natural visual systems, and the formulation of the Retinex operation that we have above is essentially due to Horn. Marr has suggested that one function of the retina is the computation of the Retinex lightness function along four channels (rod and 3 coloured cones) simultaneously, and has shown how the structure of the retina may be adapted for this purpose.

The Retinex operation is clearly well-adapted to being carried out by special parallel-processing hardware, being essentially a one-level process, independent of any particular knowledge about the scene being looked at.

(c) Texture:

Less effort has been devoted in AI to outdoor scenes than to toy scenes, probably because the problems are more difficult in the former. What does seem clear is that schemes for handling natural scenes which depend on detecting the edges of uniform intensity are inadequate: what are needed are mechanisms for the analysis of depth and texture. The analysis of texture is a particularly difficult problem.

First of all there can be a hierarchy of textures within a scene. Secondly, it is very difficult to arrive at satisfactory descriptors of textural features. That texture and related components are important in human vision is obvious from studies on the frequency characteristics of the visual system. Texture is a spatial-domain phenomenon, but there have been some proposals to treat it in the Fourier domain. A problem with such transformations is that they are essentially bulk descriptors: they smear out spatial information. We will treat this point further below. An ideal description of texture has to be multi-level, so that both small local features and larger ones which, say form boundaries, can be captured by it.

Campbell and others (e.g. Campbell & Robson, 1968) have demonstrated psychophysically that visual thresholds are directly related to the spatial frequency components of the stimulus field. They also showed that there are cells in the cat striate cortex which are highly selective for the spatial frequency of gratings over a wide range of frequencies. Pollen (Pollen & Lee,1971) demonstrated that the

simple cells described by Hubel and Wiesel are sensitive both to the stimulus area and to its brightness, the implication being that the data from a single simple cell cannot therefore provide a unique characterisation of a stimulus. He went on to claim that the visual cortex operates by a technique of strip integration which can be described as an operation in the Fourier domain. We take the view that the so-called 'Fourier theory of vision' is of limited utility in a visual system.

The Fourier transform model of vision is a version of the template matching paradigm, which holds that the primary task of a visual system is recognition, i.e. matching an image (of an isolated object) with templates from previous images. The templates could be portions of images or quantities derived from images, such as moments or Fourier coefficients. The set of problems motivating this paradigm is classification of isolated, two-dimensional forms among a small set of possibilities: character recognition is typical. We have contended that template matching is inadequate for the visual requirements of a human; unfortunately, the Fourier model is not very effective even for template matching. A model for template matching is: portions of previous images are matched against similar portions of the current image. But what part of the original image should be taken as a template? It appears that the visual system is a system capable of segmentation rather than a template matching mechanism. The templates could be supplied by revelation, or inferred from another (non template matching) facility such as motion. In the template-matching paradigm, the chief difficulty is the computational effort in matching templates to scenes involving rotation, dilation and translation of objects. But

these are only the simplest problems: we contend that real world visual problems involve articulation, obscuration, and judgments of similarity of objects which are arbitrarily different according to non-trivial metrics of template matching. The class of cups is not identified by any unique global shape or by enumeration, but by being open container (capable of holding liquid) of a certain volume. Thus, similarity depends on a description of three-dimensional form; similarity judgments often also depend on facilities segmentation and local description. Even within the template matching paradigm, the Fourier transform model has serious difficulties. The Fourier transform is equivalent to the original image; it is useful only if there are great simplifications in the frequency domain. The supposed advantage of the Fourier transform is translational invariance. However, the transform is translationally invariant only for periodic functions. Objects are finite and images are finite. In this case, the transforms are position dependent. Another difficulty is that if there are several objects in a scene, the Fourier transform depends on all of them. The spatial template matching at least allows some localization to the template.

Experiments show that linear systems analysis is useful for describing the response of the visual system to various stimuli. Further, there appear to be separate channels with about one octave width. The alternative, that the system acts as a single linear filter, seems like a straw man set up only to be knocked down, in view of the wide range of tasks facing the visual system. Experiments show that the detection threshold for square waves can be predicted

adequately from the component of the fundamental frequency. Further, square waves are distinguishable from sine waves only at contrasts such that the third harmonic is above threshold. These results do not discriminate against an edge detection mechanism, however, since the ratios of sensitivities for square wave to sine wave would be very similar, and whatever means is used to discriminate between the two stimuli must give similar results, i.e. must be dominated by the third harmonic. Any edge detector of fixed size must have a frequency response with approximately one octave width. If for example, the spatial weighting is unity across the detector, the response is zero for sine waves with period half the width and the response is half for sine waves with period twice the width.

In the light of these misgivings, we must take some pains to separate out those features of a Fourier-domain description which are useful and those other conjectures which have little support, either theoretical or experimental. An attempt to define useful textural descriptors is an simultaneously an attempt to deal with problems of pattern grouping and proximity in an n-dimensional feature space. Experimental studies of proximity have been carried out by Julesz (Julesz,1971), by Shepard (Shepard,1964), Kruskal(Kruskal,1964) and others. Shepard and his colleagues have developed some powerful scaling techniques for extracting important stimulus dimensions from data about sensory judgements, and Julesz applied them to judgements of similarity of visual textures; his finding was that the most dominant features were brightness(contrast) and orientation. Bajcsy's

(Bajcsy,1972) work attempted to use some techniques in the Fourier domain to describe such important features, and derived a way of mapping the information from the power spectrum of a scene into textural properties in the spatial domain. Obviously the power spectrum of a scene is invariant under translation (if one ignores windowing effects stemming from the finite character of images), but not under rotation, so it provides a way of specifying directionality. The phase spectrum on the other hand can be used to specify position in the scene. She was able to show how such a powerful set of descriptors for texture could be constructed and used in an algorithm for region growing. The decision problems are quite difficult, since what is essentially involved in expanding regions is a large number of judgments about proximity in n-space. The sheaf representation that she used allows one to formalise the transition from local judgments to more global structures, the whole process being imbedded in a hypothesis making/verification paradigm. Her program was able to extract significant textural information from outdoor scenes involving trees, water, grass and so on, and use this information to segment the scene in natural ways. She also devoted some thought to the significance of texture gradients in the estimation of depth, along the lines suggested by Gibson (op.cit)

(d) Depth:

The measurement of depth is another good candidate for a process which may essentially be carried out in parallel with little information other than the raw intensity data from two separate viewpoints a small angle apart. The pioneering

work of Julesz(op.cit.)has shown clearly that stereoscopic depth information may be gained in the absence of monocular features by a bulk correlation between images from the two eyes. Of course, in normal vision, many other cues interact with this straightforward computation (for example, monocular features, movement parallax, perceived size and interposition). Global features, and assumptions concerning the overall structure of the scene can be expected to help in removing local ambiguities. Blakemore, in an extremely interesting study (Blakemore,1970) has shown that the cat's visual cortex has a joint feature/depth representation, in that all orientation-specific columns of binocularly-driven cells are of either a constant-depth type, viewing a thin sheet of visual space, a few degrees wide, at a given distance, or of a constant-direction type viewing a cylinder of visual space directed towards the inter-ocular axis. The binocularly activated cells were optimally stimulated by disparities of about 0.3 deg. horizontally and 0.7 deg. vertically, and the receptive field sizes were comparable to monocularly activated cells. Adjoining columns of depth-specific cells differ by about 0.6 deg. while constant-direction columns differ by about 4 deg. So there are probably about 500 constant-depth and 300 constant-direction columns covering the entire visual field.

A program has been written at the Stanford AI Laboratory (Pingle & Thomas,1974) which carries out a feature-driven bulk-correlation able to achieve a resolution of about 1mm. at one metre. It is envisioned that this program will perform a crude 3-dimensional segmentation of a visual scene as a preface to further analysis.

4. THE INTERMEDIATE LEVEL.

In this section we move on to consider the possible behaviour of a vision system which has been provided with the kinds of information which we discussed above. In particular, we will be concerned with separating the segments of objects from the background and from each other, and with their association into coherent three-dimensional bodies. The stress will be on the wealth of possible partitionings which are possible for any reasonably interesting scene, and with the enormous computational problems which arise from a failure to introduce an adequate and systematic semantics for pictures, where by 'semantics' here we mean a set of interpretative rules consonant with the large body of facts, both intuitive and intellectual, which human observers bring to bear upon their percepts.

The psychological literature on perception of scenes is rather confused, and is really concerned only with perception of characters (alone and in context) and line drawings of 3-dimensional objects; only a very few studies have ever been done on perception of real world scenes containing large numbers of objects against a complex background (e.g.Biederman,1972; Biederman,Glass & Stacy,1973)

Some of the earliest and most significant studies of the perception of figures were those on perception of retinally stabilised stimuli, beginning with those of Pritchard et.al (Pritchard,Heron & Hebb, 1969) and of McFarland (MacFarland, 1968). When images of line drawings are optically stabilised on the retina, so that they do not move relative to the receptor surface when the eyes

move, the figures are at first perceived normally, but soon the perceived image begins to disintegrate and eventually disappears altogether. This is obviously because of the fatigue of the receptors. What is interesting is that the figures do not disappear wholistically, but rather the parts disappear independently - the first things to go are vertices, followed by line segments. One might therefore imagine that vertices and line segments form the primitive structures in perception. One might go on to suggest that a good way of studying the perception of line drawings would be to carry out eye-tracking studies, and this was indeed done by Kaufman and Richards (Kaufman & Richards,1969). They did show that acute-angled vertices are better attractants of visual fixation than obtuse ones. However they also showed that in figures subtending less than 10 deg. of visual angle, very little scanning if any is done. Instead the eyes tend to fixate rigidly upon some point within the envelope of the figure. To be more exact, fixation was predominantly upon that point which would form the centre of gravity of the object, *were it three-dimensional.* Even at this primitive stage within the visual process, prejudices about the interpretation of the line drawing are already affecting the perceptual mechanism.

Other studies have examined the notion that three dimensionality and simplicity are two organisational principles which are somehow inherent in the visual process. These two principles are but facets of the old Gestalt *Praegnanz* law, which said that scenes are perceived according to the simplest interpretation which is compatible with the sensory data. Hochberg and McAlister (Hochberg & McAlister,1953) conducted some experiments on the perception of line drawings

representing a cube in various orientations (due originally to Kopfermann). The clear cut result of their study was that the figure possessing the best symmetry as a *two* dimensional pattern was *least* often seen as a *cube* (drawing (a) in Figure 1.) This leads to a connection with the rules of symmetry much discussed by the Gestalt psychologists. Consider the kinds of organisational groupings that are going on in figure 2 (after Attneave,1968). The case of the clusters of triangles is a very interesting one - considered in isolation, any one of these triangles can appear to point in one of three directions. But when considered as clusters , all the triangles in a group appear to point in the same direction (even though this dominant direction changes from time to time). The conclusion from phenomena such as these is that the perceptual system cannot simultaneously employ more than one axis of symmetry. Attneave (op. cit.) pointed out that when the triangles are viewed from an acute angle, they optically become isosceles, but perceptually they keep their ambiguity of direction. He suggested that the reason for this is that competition between the various axes of symmetry is going on not at the level of the retinal projection but rather within an internal model of 3-dimensional space. He further suggested that this internal model may consist of a Cartesian framework, wherein groups of figures such as those above may be described in two ways:

(a) in terms of local (figural axes)

(b) in terms of the difference between the figural axes and the general axes of the visual field.

The Praegnanz principle would then dictate that the description chosen in a

particular case would be that which is simplest in terms of the orientation axes. Attneave has done some other studies suggesting that we do indeed have some representation of three dimensionality in our perceptual world models. These involved showing that the apparent tridimensional orientation of a cubical figure (represented by a line drawing) is determined by tendencies to make the object as simple as possible. For example, a figure such as that in Fig.3 could represent any of a large number of possible parallelopipeds: however if one assumes that this figure is a slanted symmetrical cube, then quite uniquely, the perceived angles α,β and γ are all equal and right-angles. This homogeneity of angles conveys a simplicity to the figure and therefore it should, according to the minimal complexity principle, be perceived usually as a cube (Attneave&Frost,1969)

Another interesting example of a possible organisational principle operating in the perception of polyhedra has been suggested by Perkins (Perkins,1968) in his analysis of the possible configurations that may be assumed at cubical corners. Some combinations of three lines meeting at a vertex look like corners of cubes,

Figure 4 about here

while others do not. Why is that? Obviously this phenomenon is governed to some extent by context, but there is also an internal principle operating. It seems a reasonable empirical conclusion to say that 'three lines meeting at a vertex form an acceptable representation of a two-faced cubical corner if and only if it contains 2 angles less than or equal to 90 deg. whose sum is greater than

or equal to 90 deg.'. We can then calculate which three line configurations form viable corners.

This little study by Perkins is a foretaste of the things we will talk about next: attempts by designers of artificial vision systems to derive systematic rules for segmenting scenes, particularly those containing configurations of regular polyhedra. We hope that corollaries with the psychological evidence that has just been presented will become evident.

The earliest work in this field was done by Roberts (Roberts,1965) who concentrated on segmenting and recognising scenes containing only blocks of various regular shapes.The input was in the form of perfect (i.e. noiseless) line drawings.

Adolfo Guzman's work (Guzman,1968) was a considerable advance, since it was the first attempt to impart some meaning to the parts of pictures (meaning in terms of the relationships between parts). This approach mitigated the problems with the template matching approach which was underlying in Roberts' program. Guzman's program is two-pass: on the first pass it gathers local evidence about vertices ; the second pass attempts to use this evidence to achieve a plausible segmentation of the scene. Again the input was a noiseless line drawing. The local evidence was based on certain heuristics concerning possible configurations of plane regions at vertices.

Figure 5 about here

Each type of possible vertex contributes information about the relations between neighbouring regions. The "arrow" link provides evidence for some connection between the two regions on either side of the shaft. The "fork" configuration provides evidence for linkage between the three contiguous regions. The "TEE" junction offers evidence of occlusion. Internally, the relationships between faces is represented symbolically by associated lists. At the lowest level, any two regions with a link between them are considered to be part of the same body. This can lead to trouble when fortuitous coincidences occur: the answer to this liberality of region association is to require that there be two links between the regions before they are considered to belong to each other. Some inhibitory heuristics were also used: for instance, there had to be compatibility of interpretation between vertices belonging to the two ends of a given line. Guzman's program worked quite well on scenes which contained large numbers of bodies in arbitrary configurations, but was bothered by holes and other non-convexities. Further, and more important, it was very poor at segmenting pictures which were missing data, e.g. line-segments. Falk (Falk,1970) approached this problem by using Guzman type heuristics for associating *edges* rather than regions. He was the able to use the verification methods we discussed earlier in connection with edge-following to predict the position of lines in the scene: here is a classical example of the kind of methodology that we have been advocating - a (fairly) high-level program concerned with analysing regions into bodies is able to induce the lowest levels of the vision process to look again for edges in sensitive and

crucial locations. For analysing real pictures, this method is quite superior to Guzman's.

Huffman (Huffman,1971) and Clowes have, independently, investigated in a systematic way the possible configurations of trihedral vertices seen from a position of non-singular perspective, classifying edges as *Convex*, *Concave* and *Obscuring*.

Figure 6 about here

Such a labelling system immediately imposes quite stringent restrictions on the possible interpretations of the vertices in this cube. They were able to show that only six L vertex labellings and three each of fork and arrow labellings were possible given the above kinds of constraints. In his paper on Impossible Objects (op.cit), Huffman points out that there is a similarity between the problem of attaching realisable labels to an arbitrary picture and that of parsing a sentence in a language: he is able to suggest that the reason impossible objects are so difficult to understand is that they are the embodiments of illegal parsings . Waltz (Waltz,1972),in his recent thesis, developed these ideas further: he expanded the set of possible labellings to take into account shadows and other vagaries of illumination, and introduced a filtering mechanism involving resolution of conflicting information from neighbouring edges as well as a few heuristics concerning legal lighting situations, etc. which allowed his program to converge rapidly onto a plausible interpretation, rather than carrying out a depth-first search of all the possible Huffman labellings.

5. A GATHERING TOGETHER.

This section will attempt to draw general conclusions from the detailed points that have been made above, will lay out some suggestions for description schemes for real scenes, and will go on to contend that these arguments have revelance for the contemporary controversy in cognitive psychology about the nature of mental representation.

Given the initial point of view that the major perceptual problem is that of description rather than recognition, several conceptual difficulties arise which do not seem to have received the attention we think they deserve, either in the psychological literature or in designs for artificial vision systems. They are, firstly, that of attention : the problem of deciding at what level of analysis (in the sense of Section 2) to approach a perceptual task; and, secondly, that of arriving at representational formalisms which capture in some smooth way the imaginal and symbolic aspects of human description schemata.

> *'By "description" we do not usually mean verbal descriptions; we mean an abstract data structure in which are represented features, relations, functions, reference to other processes and other information. Besides representing things and relations between things, descriptions often contain information about the relative importance of features to one another, e.g. which features are to be regarded as essential and which are merely ornamental.' (Minsky & Papert,1972)*

The attention problem is intimately bound up with the particular goals of the system at any given moment. For example, in driving a car, one's attention is normally devoted to the road directly in front, but any unexpected motion in the peripheral

visual field receives immediate notice. In AI, this behaviour has been captured by the notion of a 'demon', i.e. a high-priority interrupt process which, when activated by a specific event (e.g. motion) can stop whichever computation is presently in progress and cause some other process to be run to deal with the interrupt. This would seem to be easily implementable in a natural vision system where problems of scheduling are reduced by the inherent parallelism. Most attempts at the development of artificial vision systems have devoted their efforts to extremely detailed analysis of the whole visual field, involving the actual recognition of every visible object, usually by matching against some proto-type which has been learned and stored in memory. This prototype-matching matching approach, which is in some ways a generalization of the template-matching techniques which we scorned earlier, also fails to meet our second criterion of adequacy, in that it is not easily expanded to a smooth representational formalism: it is overly concerned with localized features, at the expense of understanding larger-scale characteristics of a visual scene.

We will discuss three approaches which begin to escape these apparent defects: the first, described by Binford (Nevatia&Binford,1973), involves a generalised volume description of curved objects; the second, due to Winston (Winston,1970) is the beginning of an attempt to describe the relations between objects in a scene; and the third is Minsky's recent suggestion concerning the notion of frames (Minsky,1974)

A representation useful for the work of Binford et al. with complex objects was chosen to satisfy design criteria which are relevant to human perception. The

representation must be generative, that is, a rich set of shapes should be conveniently generated from local primitives. For that to be effective, the representation should have a segmentation into parts, which themselves may be formed of other parts. For the part/whole segmentation to be effective, the primitive parts must be naturally defined and computationally adequate. In this case it was decided to define primitive parts by continuity. This implies that the primitives are volume primitives and not surface primitives, since surfaces are discontinuous for objects with what we conveniently think of as a single part (e.g. a block).

The representations depend upon segmentation to describe complex objects in terms of parts. The representations are graph-structured; nodes correspond to parts, while arcs correspond to relations between parts. Parts may be compound parts, i.e. graphs of the same form. Relations include relative position and orientation, degrees of freedom, symmetry and any special knowledge available. The topological operations of cutting and pasting are used in joining parts. Normally, we think of holes as made by cuts and protuberances as made by pasting.

Primitive parts are arm-like, described as "generalized local cones". These parts are described formally in terms of "generalized local translational invariance", appropriate to parts whose cross sections change slowly along some space curve. A general cylinder is formed by an arbitrary cross section translated along a straight line. A cone has a linear variation of cross section along this axis. If the scale of the cross section is varied smoothly along the axis, we have what might

be called a local cone. If the cross section is allowed to vary by distortion or rotation, and the axis is allowed to be a space curve, then we have "generalized local cones". These are the volumes swept out by taking an arbitrary cross section and translating it along a space curve, meanwhile varying the cross section while holding the cross section normal to the path.

Figure 7 about here

Cross sections are represented in the same fashion in two dimensions as objects are in three dimensions. Again, part/whole segmentation is crucial. It is required that parts be defined by continuity. Thus the parts must be area parts, not curve parts, since curves are discontinuous for plane figures which we normally think of as having a single part. The primitives are described by a cross section (one-dimensional) varying smoothly along a plane curve.

Figure 8 about here

Each element has a local coordinate system. Each joint contains the transformation necessary to go from the coordinate system of one element to that of the other. This process of segmentation allows non-unique representations, and permits us a choice of simple representations; we can regard the non-uniqueness as an advantage in the light of our comments above about the attentional problem.

Figure 9 about here

The basic philosophy of Winston's work is two-fold:

(a) that learning is a process of constructing compact descriptions of the meaningful relationships between objects in the world; and

(b) that such description construction is not to be achieved by a statistical learning paradigm, but rather by judicious teaching, i.e. the choice by a teacher of examples which reveal the crucial features of a relation-description. Very importantly, it may involve exposing the learner to what Winston has called 'near-misses': situations which differ from the one another in some singularly crucial way.

Figure 10 about here

Figure 10 shows the kind of data structure that Winston's program constructs to describe the spatial relationships involved in a simple pedestal. One of the most crucial relations in this construction is that of *support* - a feature which is absent from the 'near-miss' examples. The role of the teacher in this case is to reinforce by means of such examples the crucial relationships which obtain in the scene. Obviously there can be a hierarchy of relations which can be exhibited by such examples, some of which are more crucial than others.

In this work, and in that of Binford, we think that there are the beginnings of a viable and fruitful theory of scene descriptions based on structural descriptions in which the topology is manifest. Developmental studies of human perception, particularly by Piaget and his colleagues (Piaget&Inhelder,1967) have of course revealed that the ability to construct toplogical descriptions occurs earlier than a more metric and geometric ability. Minsky(*op.cit.*) in an interesting paper, has

developed the notion of a 'frame' which is essentially a generalization of Winston's scheme particularly adapted to revealing at its top level those features of a situation which are believed to be most important:

> *'Seeing is based on the use of Frame-Systems. A frame-system is a data-structure for representing a stereotype situation - like being in a certain kind of room. Attached to this structure are several kinds of information. Some of this information is about what one expects to happen next. Some is about what to do if expectations are not confirmed.*
>
> *Collections of related frames are linked together into Frame-systems. The effects of important actions are mirrored by transformations between the frames of a system.' (Minsky,1974)*

We do think however that the essentially non-metrical character of Minsky's frames is somewhat unsatisfactory. His argument seems based on the belief that 'people do not seem to have much metrical ability vis-a-vis three dimensional imagery' (Minsky, op.cit.); this does not fit well with the studies of Attneave that we discussed above, which seem to demonstrate the contrary.

This difficulty leads us naturally to discuss the relationship between symbolic and pictorial methods of representation, and in particular the nature of mental imagery. Of late, there has been considerable controversy in the literature (represented by Minsky's paper discussed above, and by an elegant discussion due to Pylyshyn (Pylyshyn,1972)). We will not, in this essay, devote detailed discussion to this problem: a paper is forthcoming by one of us (AJT) which argues, *contra* Pylyshyn, that mental imagery is a valid phenomenological concept, and that while it may be true that the general nature of memory structure is that of a symbolic network, it

does not therby follow that either the subjective experience of imagery or the considerable body of structural investigation, e.g. by Shepard (Shepard,1971) and especially by Cooper (Cooper,1973) on mental imagery tasks involving random three-dimensional forms, is thereby to be faulted. The argument is basically that the symbolic/ pictorial distinction is a functional and computational one, and that some form of the dual-code hypothesis advocated, for example, by Posner (Posner,1972) is a likely beginning of a satisfactory theory. The major question is whether it is correct to think of perception as a combination of a distal stimulus and a mental image thereof, and conversely whether the perceptual machinery is active during pictorlal imagining. Advocates of a dual-code type of hypothesis,(e.g. Bower,1972) have argued in favour of a commonality of generation of both pictorial and linguistic structures. Powerful and interesting evidence for a role of perceptual concepts in linguistic operations comes from studies of acquisition of language by children. Postal's Universal Semantic Primitives Hypothesis (Postal,1966) makes the claim that linguistic primitives must reflect closely the primitive relationships extant in the organism's world, while Bierwisch (Bierwisch,1969) has made the point that not only must a child have such a set of primitives, but he must also learn to recognise them for what they are. E.Clark (Clark,1971) has suggested that at an early stage a child who is just beginning to use words does not know their full meaning, but rather identifies them initially with only a few features of the meaning, which are criterial for its use of the words. From our point of view here, the crux of her hypothesis lies in the nature of such features:

> *'...the first semantic features that a child uses are liable to be derived from the encoding of his percepts;... at a later stage, as the child learns more about the structure of his language as a whole, he will learn which percept-derived features play a particular linguistic role, and which are relatively redundant'(Op.cit.)*

For example studies on over-extension of relational terms (e.g. the use of the term 'dog' to name all four-legged animals, independently of size, &c.) are particularly interesting: when two words are closely related, e.g. *more/less, same/different, before/after*, the tendency in confusion of use by a child is towards the use of the unmarked member of the pair in place of both meanings. In the st dies by Piaget and others on over-extended use of synonyms eg. boy/brother, there is a suggestion that the tendency is to use such words synonymously until the relevant discriminating features are learned and added to the lexical entry. H.Clark (H.Clark,1971) has carried this way of thought as far as to suggest that an isomorphism exists between the linguistic and perceptual domains. He proposes that asymmetries seen in the use of pairs of polar adjectives e.g. big/small, tall/short, in front/behind are due to analogous asymmetries which exist in the visual field. It is as if there existed reference points within this field, about which directions tend to be defined as either positive or negative, depending on their perceptual predominance.

6. CONCLUDING REMARKS.

We hope that, throughout the preceeding sections, the reader has been able to trace the two skeins of our argument:

(a) that a vision system, whether natural or artificial, cannot function in the absence of a two-way flow of information to and from almost every stage of analysis; and

(b) that the crucial problem in understanding or designing such systems must be the formulation of a representational formalism which captures both perceptual and non-perceptual information, and allows a smooth transition between them. We have sketched what we take to be the beginnings of such a formalism, but we find that we have to admit that even these beginnings are at present quite incomplete and unsatisfactory.

Our hope is that this paper will have gone some way towards reducing any difficulty due to terminological and conceptual differences which exist between AI and psychology, and which have inhibited a potentially fruitful discourse.

References.

AGIN, G.J.(1972) 'Representation and description of curved objects', Stanford A.I. Memo No.173

ALBERTI, L.B. 'Della Pittura', Venice (1547)

ANDERSON, J. & BOWER, G.(1973) 'Human Associative Memory', New York

ATTNEAVE, F. (1968) 'Triangles as Ambiguous Figures', American Journal of Psychology,81:447

ATTNEAVE, F. & FROST, R.(1969) 'The determination of perceived tridimensional orientation by minimum criteria', Perception and Psychophysics, 6:391

BAJCSY, R. (1972) 'Computer Identification of Textured Visual Scenes', Stanford AI Memo.#180

BIEDERMAN,I. (1972) 'Perceiving Real world scenes', Science,177:77

BIEDERMAN, I.,GLASS,A.L.& STACY,E.W.(1973) 'Searching for Objects in Real-world Scenes', Journal of Experimental Psychology, 97:22

BIERWISCH, M.(1969) 'On Certain Problems of Semantic Representations', Foundations of Language, 5:153

BLAKEMORE, C. (1970) 'The Representation of 3-dimensional Visual Space within the Cat's Visual Cortex' Journal of Physiology(London), 209:155

BORING,E.G.(1942) 'Sensation and Perception in the History of Experimental Psychology', New York and London.

BOWER, G.H.(1972) 'Mental Imagery and Associative Learning' in L.Gregg(ed.) Cognition in Learning and Memory', Wiley,NY.

BRINDLEY, G.S.& MERTON,P.A.(1960) 'The Absence of Position Sense in the Human Eye', Journal of Physiology(London), 153:127

CAMPBELL, F.& ROBSON,J.G. (1968) 'Application of Fourier Analysis to the Visibility of Gratings', Journal of Psychology, 197:551

CLARK, E.(1971) 'What's in a Word? On the Child's Acquisition of Semantics in his First Language', in Proceedings of Buffalo Conference on Developmental Psycholinguistics

CLARK, H.H.(1971) 'Space, Time, Semantics and the Child' (Ibid.)

COOPER, L.A.(1973) 'Chronological Studies on the Rotation of Mental Images' Unpublished Ph.D. dissertation, Stanford University

FALK, G.(1970) 'Computer Interpretation of Imperfect Line Data as a three-dimensional scene', Stanford AI Memo. 132

GIBSON, J.J.(1950)'The Perception of the Visual World', Boston.

GOMBRICH,E H.(1960) 'Art and Illusion', Princeton.

GREGORY, R.(1970) 'The Intelligent Eye', New York

GUZMAN, A.(1968) 'Computer Recognition of Three-dimensional objects', MIT Project MAC TR.59.

HELMHOLTZ, H.VON(1963)'Handbook of Physiological Optics', Dover Reprint.

HOCHBERG,J. & McALISTER,E. (1953) 'A Quantitative Approach to Figural Goodness', Journal of Experimental Psychology, 46:361

HOFFMAN, K.P. (1973) 'Conduction velocity in pathways from retina to superior colliculus in the cat: a corelation with receptive field properties', JourNal of Neurophysiology, 36:409.

HOFFMAN, K.P. & STONE, J (1971) 'Conduction velocity of afferents to cat visual cortex: a correlation with cortical receptive field properties', Brain Research, 32:460.

HORN,B.K.P.(1974) 'On Lightness' MIT A.I. Laboratory Memo.

HORN, B.K.P. & BINFORD,T.O.(1973) 'The Binford-Horn Line-finder', MIT Artificial Intelligence Laboratory, Memo.285

HORN,G. & HILL,R.M.(1969) 'Modification of receptive fields of cells in the visual cortex occurring spontaneously and associated with body tilt' Nature(London), 221:186

HUBEL,D & WIESEL,T.N.(1968) 'Receptive Fields and functional architecture of monkey striate cortex', Journal of Physiology(London),195:215

HUECKEL,M.(1971) 'An operator which locates edges in digitised pictures', Journal of the Association of Computing Machinery, 18:118

HUFFMAN, D.(1971) 'Impossible Objects as Nonsense Sentences', in Michie(ed.) Machine Intelligence,6. Edinburgh.

JULESZ, B.(1971) 'Foundations of Cyclopean Perception', Chicago

KAUFMAN,L & Richards,W.(1969) 'Spontaneous Fixation Tendencies for Visual Forms' Perception and Psychophysics, 5:85

KRUSKAL, J.B.(1964) 'Non-metric multidimensional scaling: a numerical method' Psychometrika,29:115

LAND, E.& McCANN,J.J.(1971) 'Lightness and Retinex Theory', Journal of the Optical Society of America,61:1

McFARLAND,J.H.(1968) 'Parts of Perceived Visual Form: new evidence', Perception and Psychophysics, 3:118

MARR, D.(1974) 'An Essay on the Primate Retina', MIT Artificial Intelligence Laboratory, Memo 296.

MICHIE, D.(1971) 'On Not Seeing Things', Edinburgh Machine Intelligence Laboratory, Memo.22.

MINSKY, M.(1974) 'Frame-systems: a theory of representation of knowledge' Unpublished memo, MIT AI Laboratory

MINSKY, M. and PAPERT,S.A.(1972) 'Research at the Laboratory in Vision, language and other problems of intelligence', MIT A I Laboratory Memo.252.

NEISSER,U.(1967) 'Cognitive Psychology', New York

NORMAN, D.A.(ed)(1967) 'Models of Human Memory', New York

NEVATIA, R. & BINFORD, T.O.(1973) 'Structured Descriptions of Complex Objects', in Proceedings 3rd.Int.Joint.Conf. on Artificial Intelligence, Stanford.

PERKINS, D.N.(1968) 'Cubic Corners', Quart.Progress Report 89, MIT Research Laboratory of Electronics

PERKINS, W.B. & BINFORD, T.O.(1973) 'A Corner Finder for Visual Feedback, Stanford A.I.Laboratory, Memo. 214

PIAGET,J. & INHELDER,B.(1967) 'The Child's Conception of Space', New York

PINGLE,K.K. & THOMAS,A.J.(1974) ' A Feature-driven Stereo Program', Stanford A I Memo. 248

POLLEN,D.A., LEE,J.R., & TAYLOR,J.H.'How does the Striate Cortex begin the reconstruction of the visual world?', Science XXXXX

POSNER, M.(1972) 'Coordination of Internal Codes', in W.G.Chase(ed.) Proceedings 8th. Carnegie Symposium on Cognitive Psychology

POSTAL,P.M.(1966) 'Review of Martinet "Elements of General Linguistics"' , Foundations of Language, 2:151

PRITCHARD,R.M., HERON,W. & HEBB,D.O.(1969) 'Visual Perception approached by the method of stablized images', Canadian Journal of Psychology,14:67

PYLYSHYN, Z.W.(1972) 'The Problem of Cognitive Representation', Research Bull.227, Psychology Dept., Univ. of Western Ontario.

ROBERTS, L.G.(1965) 'Machine Perception of 3 dimensional solids', in Proceedings Symposium on Optical and electro-optical processing of information.

QUILLIAN,R.(1969) 'The Teachable Language Comprehender: a simulation program and theory of language'.Communications of the Association of Computing Machinery, 12: 459

SCHANK, R.C. (1972)'Conceptual Dependency: a theory of natural language understanding', Cognitive Psychology, 3(4)

SHEPARD, R.N.(1964) 'Extracting Latent Structure from Behavioural Data' in Proceedings 1964 Symposium on Digital Computing, Bell Telephone Laboratories

SHEPARD, R.N.(1972)'Cognitive Psychology of Internal Representation', unpublished grant proposal

SPINELLI, D.N., STARR,A. & BARRET,T.W.(1968) 'Auditory spocificity in unit recordings from the cat's visual cortex', Experimental Neurology, 22:75

TENENBAUM, J.M. & PINGLE,K.K.(1971) 'An Accomodating Edge-Follower',in Proc. 2nd. Inter. Joint Conf. on Artificial Intelligence, London.

THOMAS, A.J.(1973) 'A Retinex-like program for colour identification', Stanford AI Memo (forthcoming)

WALTZ,D.L.(1972) 'Generating Semantic Descriptions from drawings with shadows', MIT AI Lab. TR-271

WINOGRAD,T.W.(1972) 'Understanding Natural Language', New York.

WINSTON,P.H.(1970) 'Learning Structural Descriptions from Examples', MIT AI Lab. TR-76

WITTGENSTEIN, L.W. (1953) 'Philosophical Investigations', Oxford.

DA VINCI,L.(1956) 'Treatise on Painting', A.P.McMahon(ed.), Princeton.

FIGURE CAPTIONS

Figure 1

Possible configurations of tri-hedral vertices. Only some are convincingly parts of a cubical body(after Kopfermann)

Figure 2

Ambiguous triangle configurations (after Attneave(1968)). The equilateral triangles appear in one of three orientations depending on the dominant symmetry axis of the whole group.

Figure 3

Constraints on the perception of a parallelopiped as a cube (after Attneave & Frost(1969))

Figure 4

Some further constraints on the perception of cubical corners(after Perkins(1968))

Figure 5

Linking regions using Guzman's scheme (after Guzman(1968))

Figure 6

Huffman labelling scheme for 3-dimensional polyhedra: plus marks a convex aedge, minus a concave edge and an arrow marks edges where only one face is visible of the two which make up the edge.(after Huffman(1971))

Figure 7

Binford Generalized Cylinder representation for a screwdriver (after Agin(1972))

Figure 8

Laser ranging data for a toy doll (from Nevatia(1974))

Figure 9

Segmentation of a toy doll along several axes, derived from the range data of Figure 9 (from Nevatia(1974))

Figure 10

A structural description of a simple pedestal(after Winston(1970))

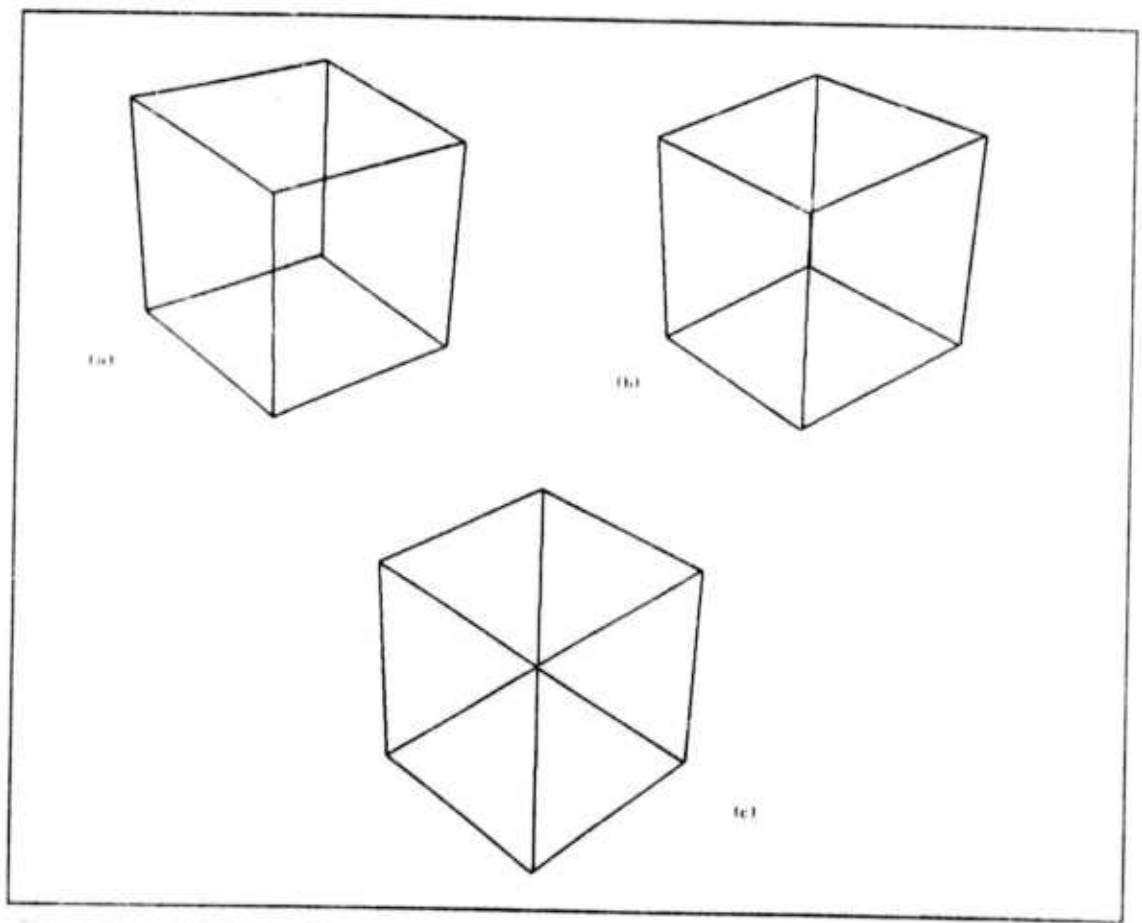

Figure 1: Kopfermann cubes

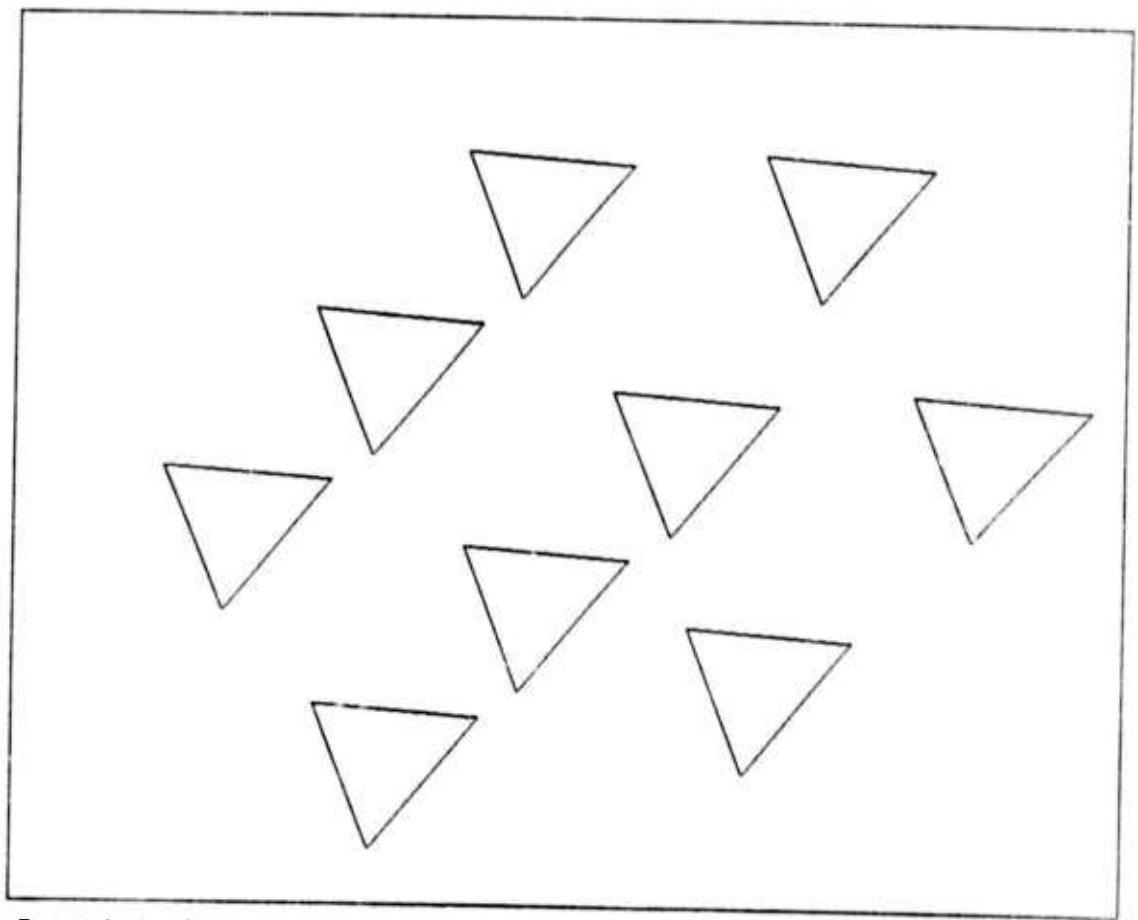

Figure 2: Ambiguous Triangles

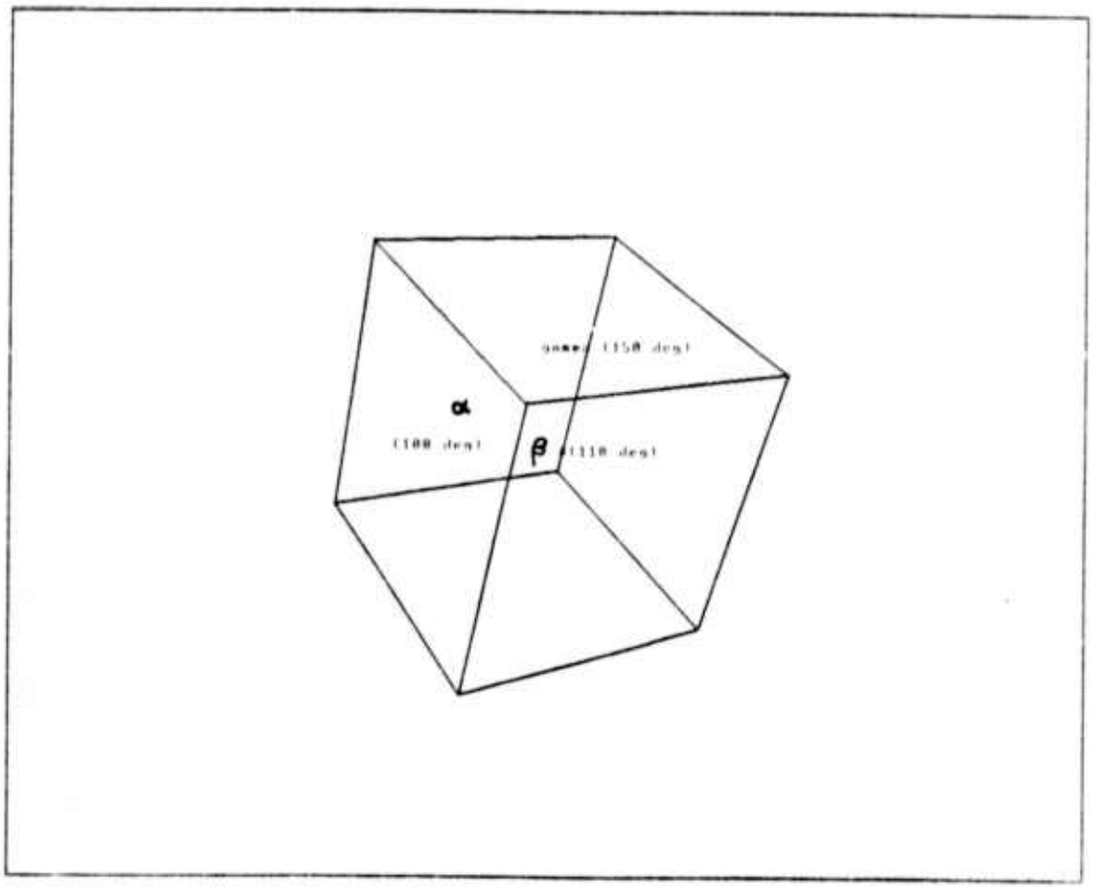

Figure 3: Permissible angles at a cubical corner

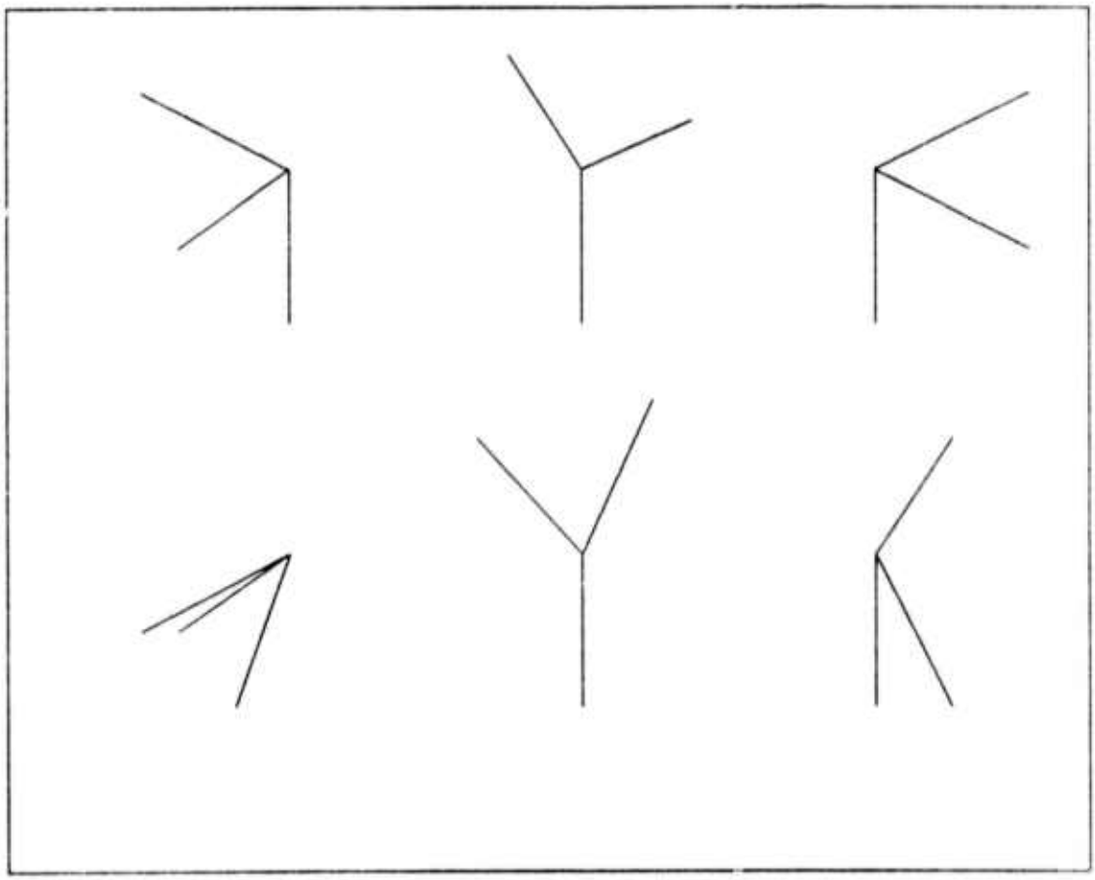

Figure 4: Possible trihedral vertices

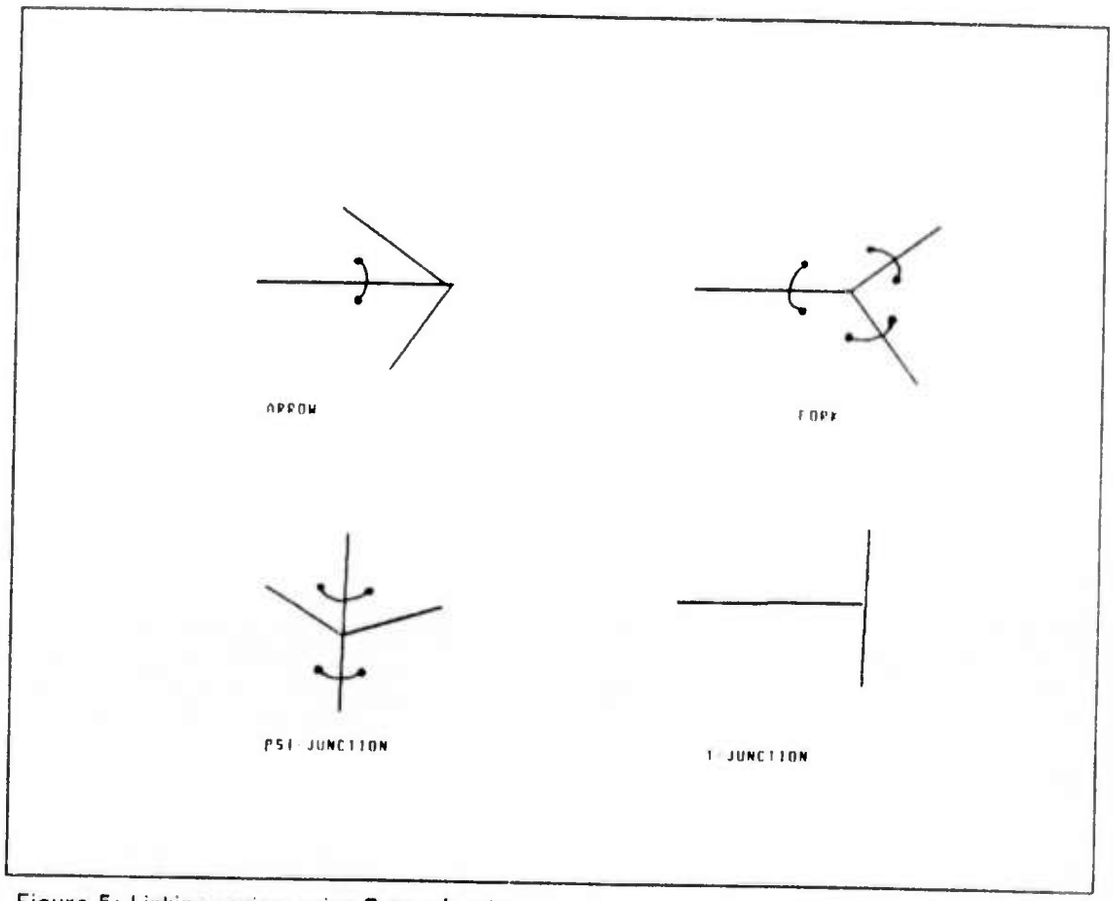

Figure 5: Linking regions using Guzman's scheme

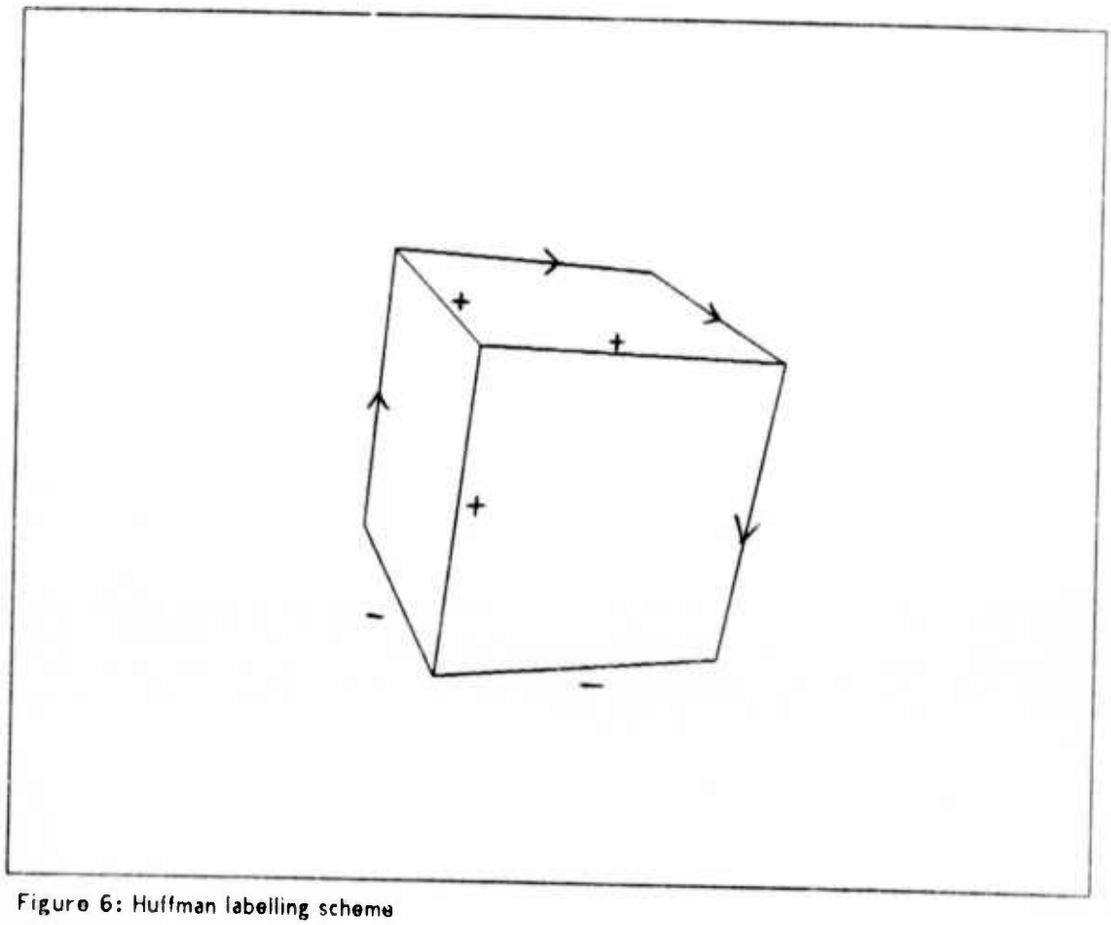

Figure 6: Huffman labelling scheme

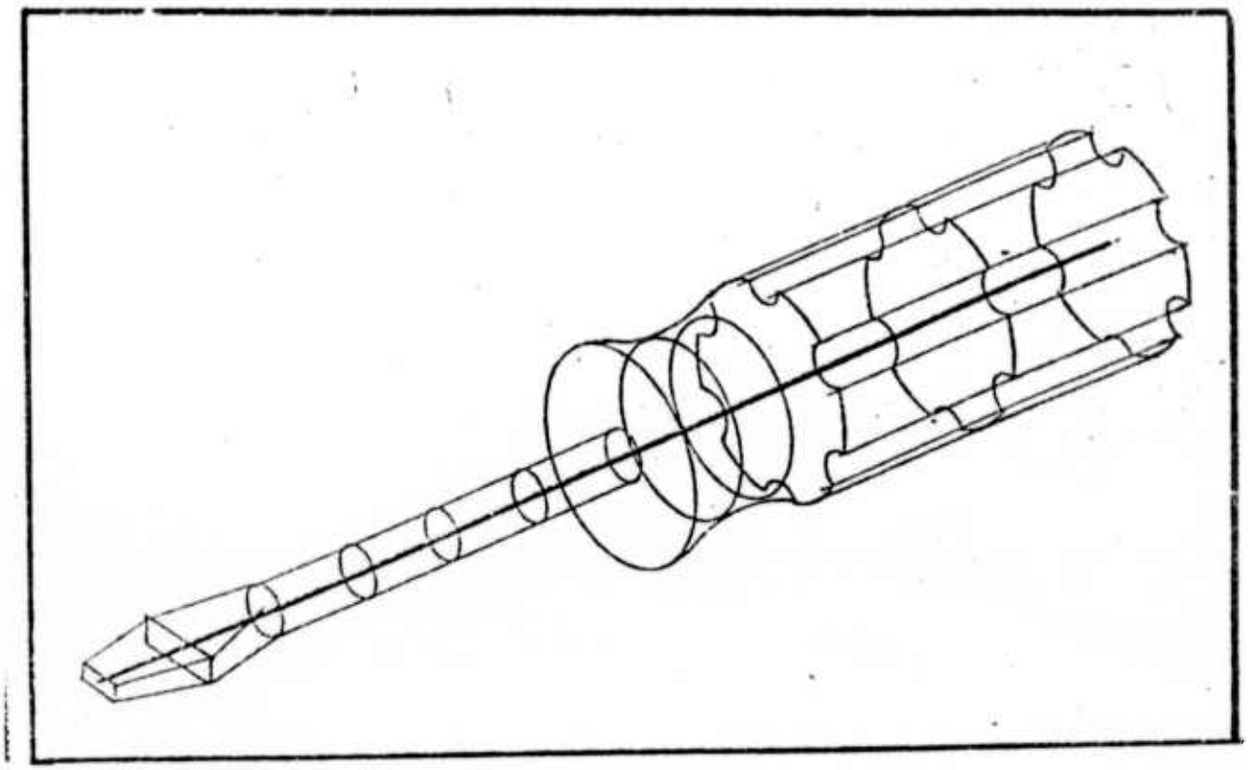

Figure 7: Generalized cylinder representation of a screwdriver

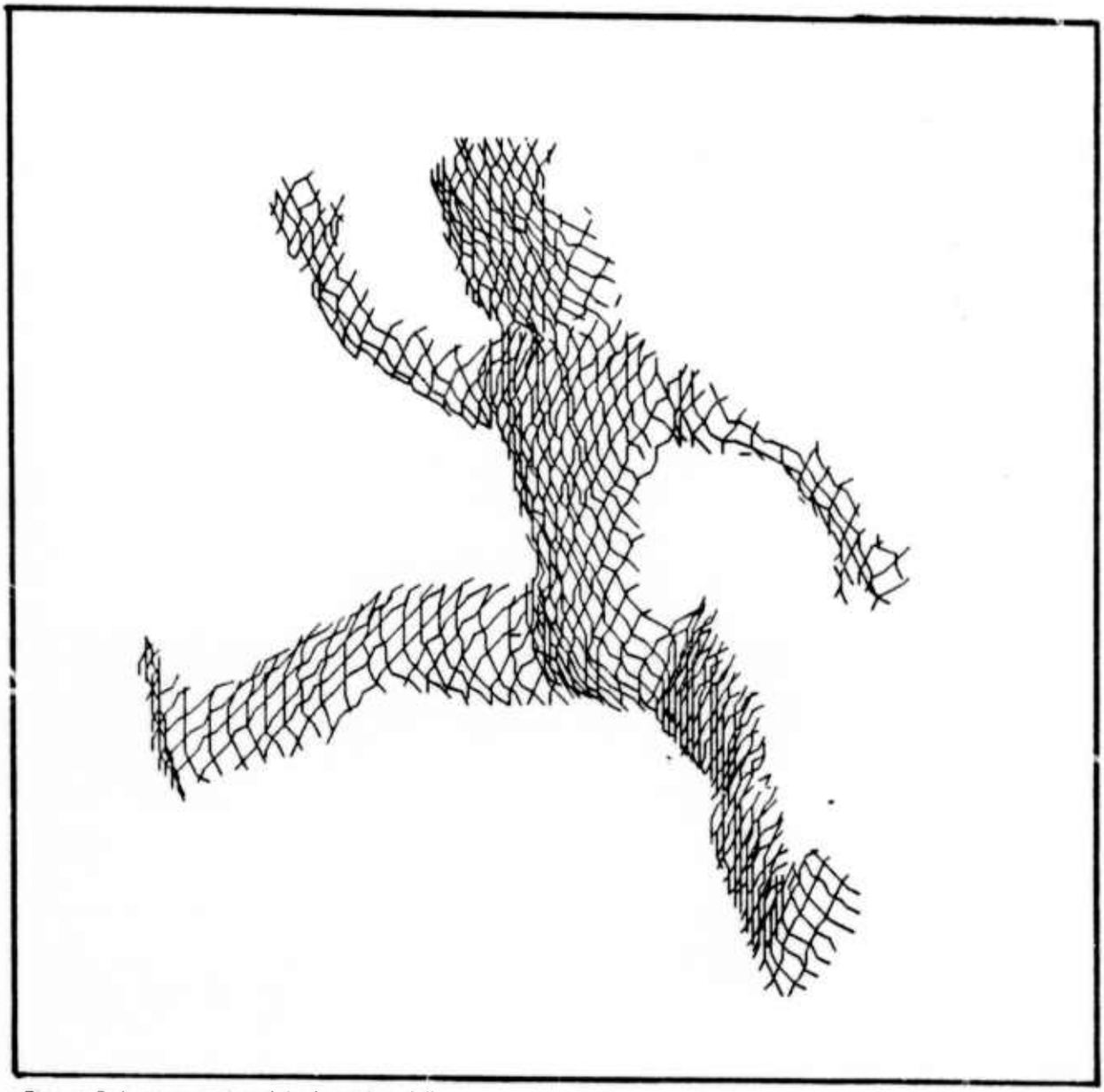

Figure 8: Laser ranging data for a toy doll

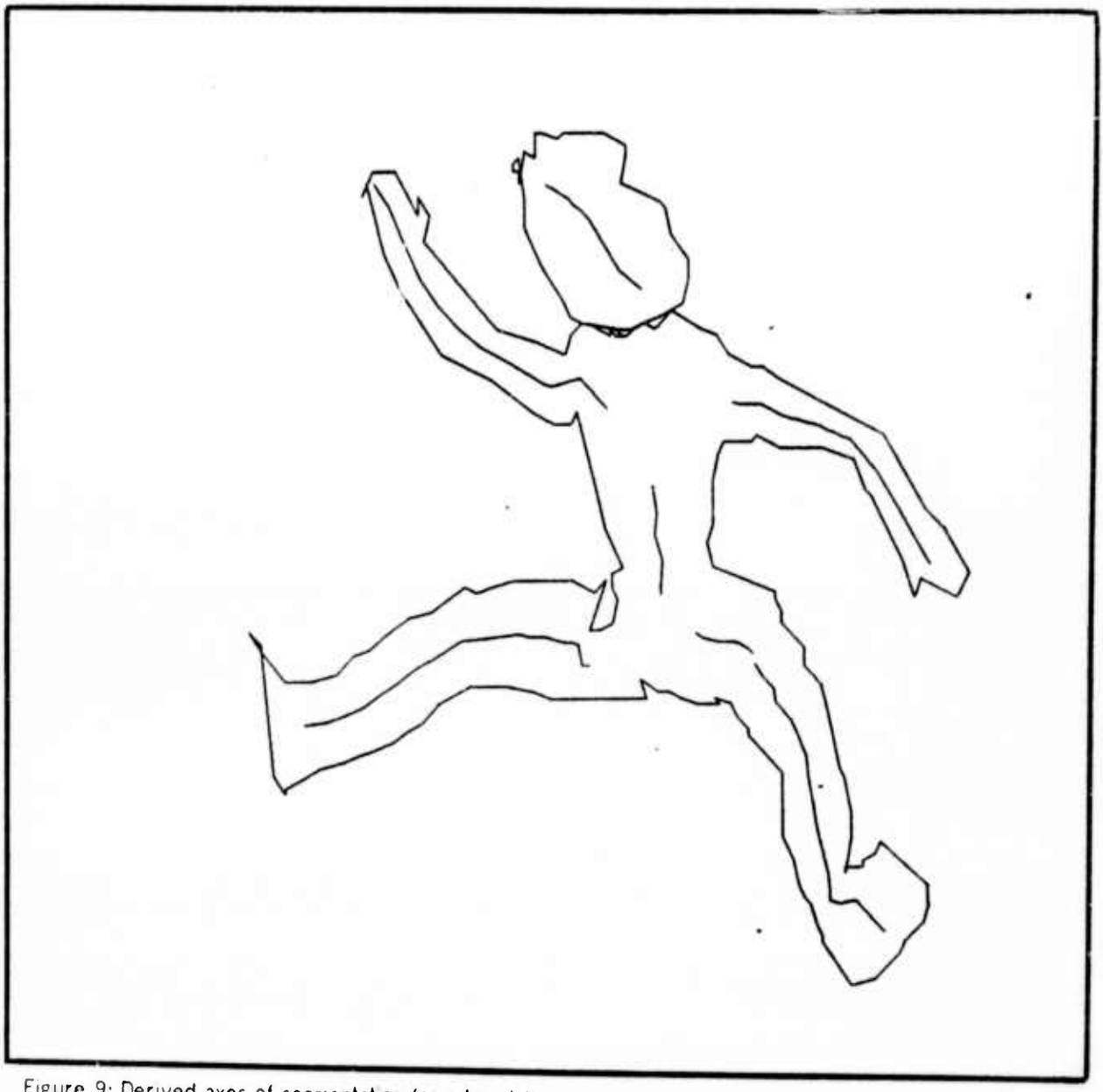

Figure 9: Derived axes of segmentation for a toy doll

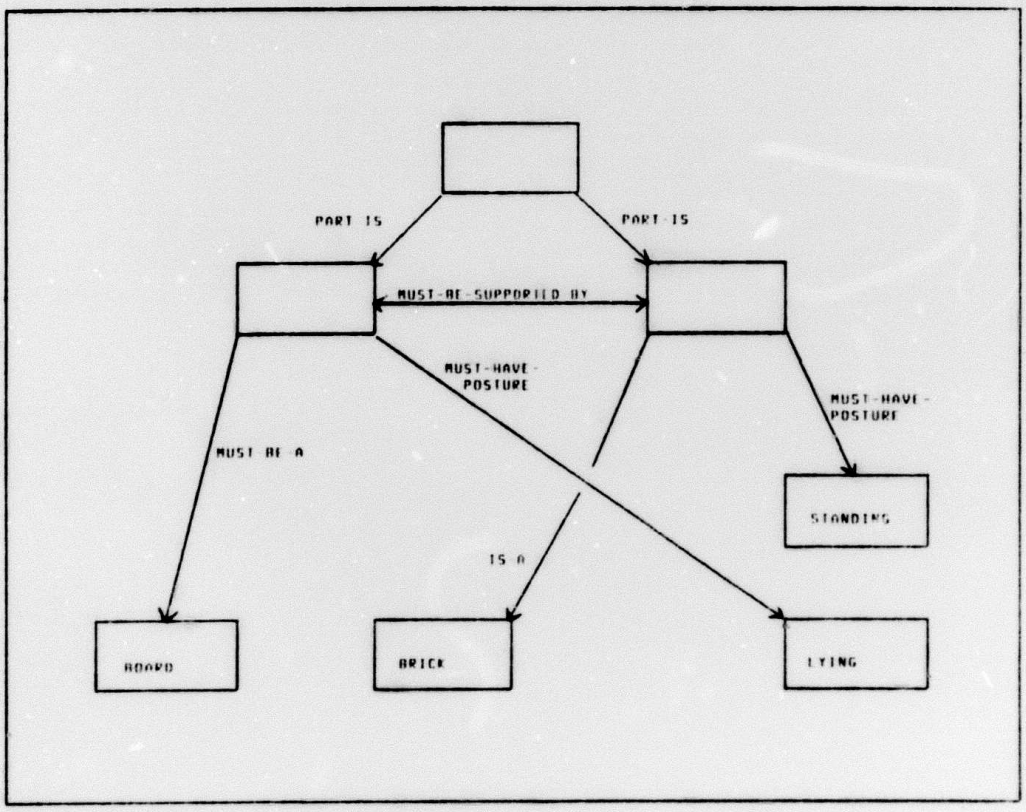

Figure 10: Winston schema for a pedestal